Title: Graph Mining

Name: Jan Ramon

Affil./Addr.: Declarative Languages and Artificial Intelligence group, K.U.Leuven, Belgium

Jan.Ramon@cs.kuleuven.be

# Graph Mining

## Synonyms

Network analysis, Learning from graph structured data.

## Definition

Graph mining is the study of how to perform data mining and machine learning on data represented with graphs. One can distinguish between on the one hand transactional graph mining, where a database of separate, independent graphs is considered (such as databases of molecules and databases of images), and on the other hand large network analysis, where a single large network is considered (such as chemical interaction networks and concept networks).

## Characteristics

### Graph-structured data

In many applications, it is natural to represent data with graphs. One can distinguish two main settings. First, in the transactional graph mining setting, databases of separate, independent graphs are considered. For example, in a molecule database, molecules are commonly represented using one vertex for every atom and one edge

for every bond between two atoms. Large, publicly available databases of chemical compounds include the NCI dataset (`http://cactus.nci.nih.gov/`) and the ZINC dataset (`http://zinc.docking.org/`).

Second, in the single (large) network setting, all data is represented in one large, connected network. Examples of such networks include the Internet, social networks, citation networks, concept networks, computer networks, chemical interaction networks, regulatory networks, socio-economic networks and encyclopedias. Sample datasets are publicly available at amongst others `http://snap.stanford.edu/data/`. In a chemical interaction network, molecules are represented by vertices connected by chemical reactions. The level of detail and the exact representation may different among datasets. For example, chemical reactions may be represented as separate nodes in the network with arcs from/to the participating compounds, or they may be implicit, in which case compounds which are involved in the same chemical reaction are just connected with an undirected edge. Next to networks of chemical compounds, it is also common to consider higher-level networks such as protein interaction networks and gene regulatory networks. For example, in gene regulatory networks nodes represent genes and arcs between nodes indicate that one gene codes for a transcription factor regulating the other gene. In comparison to the transactional setting, an important challenge in the single network setting is that one's beliefs on all data may be dependent on one another. Most traditional machine learning techniques assume that examples are drawn identically and independently (i.i.d.).

Other abstractions are sometimes preferred to graphs in order to represent similar data, such as relational databases and logic. The domains focussing on data mining using these representations are called relational data mining and inductive logic programming, respectively. Representing data with graphs has several advantages. First, the representation language is simple and therefore allows for the fast development

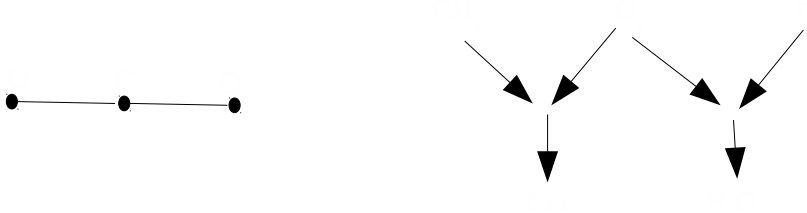

**Fig. 1.** A molecule (Carbondioxide) represented as a graph (left) and a chemical interaction network depicting the oxidation of methane and hydrogen (right)

of algorithms. Second, the representation language is expressive and adequate for the majority of applications. Finally, there is a vast literature on efficient graph algorithms. A potential disadvantage, especially in order to use algorithms implemented only for simpler graph representation, is that it may be necessary to transform the data into a simpler (but equally expressive) graph format in a preprocessing step.

## Transactional graph mining methods

Graph mining methods cover the whole range of methods from data mining and machine learning. We only list here a few examples of methods which received significant attention in the literature.

*Graph pattern mining*

Graph pattern mining methods perform pattern mining on graph-structured data, i.e. they list all patterns which satisfy some interestingness criterium such as being frequent. A frequent pattern is a pattern which is a subgraph of at least a certain fraction of the transaction graphs in the database. Well-known graph mining systems are gSpan (Yan and Han 2002) and Gaston (Nijssen and Kok 2004).

A popular strategy for the application of these systems and related ones to quantitative structure-property relationship (QSPR) modeling (i.e. the modeling of the relationship between the structure of molecules and their chemical properties) is to

first generate frequent molecular fragments, then to generate one boolean feature per pattern (with value 1 for molecules having the pattern as substructure and with value 0 for other molecules), and then to apply some suitable classification algorithm (such as a support vector machine) on these features.

*Comparing graphs*

In order to compare small graphs, such as molecular graphs, one can use graph kernels, graph metrics and maximum common subgraph operators. Kernels on molecular graphs such as presented in (De Grave and Costa 2010) can be used with any kernel-based learning method such as support vector machines and Gaussian processes. Metrics and maximum common subgraph operators can be used in instance-based learning approaches, or as features for a wide range of classification algorithms (Schietgat et al. 2010).

## Methods for analyzing large networks

*Analyzing overall network regularity*

An important starting point for many methods for analyzing large networks it the observation that large real-world networks, independently of the domain, satisfy a number of statistical regularities. For example, many networks satisfy the small world model, which informally corresponds to the fact that the number of highly connected nodes is much smaller than the number of low degree nodes. Also, many networks can be clustered in modules of nodes which are much better connected to each other than to nodes in other modules. As a consequence, much inspiration has come from random graph theory (Bollobás 2001; Durrett 2007) and spectral graph theory (Chung 1997), which study the statistical properties of such graphs. Alon (2007) discusses motifs in biological networks and the surprising deviation of frequencies of certain motifs from what one would expect if the given network were completely random.

*Predicting node properties*

Often however, in addition to network-level regularities, also a more detailed node-by-node analysis of a network is necessary. Several approaches aim at modeling properties of nodes in a network. First, in the field of statistical relational learning (Getoor and Taskar 2007) probabilistic models, are being studied which allow to reason about beliefs of the properties of individual nodes and their connections in a Bayesian network manner. Second, semi-supervised learning (Zhu and Goldberg 2009) aims at learning predictive models exploiting not only the information about the training examples but also the information about the unlabeled examples. This is especially useful in networks where nodes and their connections are known, but not the value of some target attribute.

## Cross-references

Classification

Data mining

Frequent pattern mining

Graph

Learning, kernel-based

Machine learning

Pattern mining

Support vector machine

## References

Alon U (2007) An introduction to systems biology, Chapman & Hall/CRC

Bollobás B (2001) Random graphs. Cambridge University Press

Chung F (1997) Spectral graph theory, AMS

De Grave K, Costa F (2010) Molecular graph augmentation with rings and functional groups, Journal of Chemical Information and Modeling 50(9): 1660–1668

Durrett R (2007) Random graph dynamics. Cambridge University Press.

Getoor L, Taskar B (2007) An Introduction to statistical relational learning. MIT Press

Nijssen S, Kok J (2004) A quickstart in frequent structure mining can make a difference, Proceedings of the 10th ACM SIGKDD International Conference, pp. 647–652

Schietgat L, Costa F, Ramon J, De Raedt L (2010) Effective feature construction by maximum common subgraph sampling, Machine Learning 83(2): 137–161

Zhu X, Goldberg AB (2009) Introduction to Semi-Supervised Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning 3:1-130

Yan X, Han J (2002) gSpan: Graph-based substructure pattern mining. In proceedings of the 2002 International Conference on Data Mining (ICDM'02), pp. 721–724